Data Metrics for 2020 Disclosure Avoidance
3/25/2020

The Census Bureau has been working with the data user community on a set of metrics that will allow for the evaluation of improvements through the iterative development of the 2020 Disclosure Avoidance System (DAS). This document provides information related to this effort.

We welcome feedback and questions on this document. Please submit feedback on these set of metrics by Friday, April 24, 2020 to: dcmd.2010.demonstration.data.products@census.gov.

Background

The Census Bureau is developing a new method of disclosure avoidance for the 2020 Census to protect the privacy of respondents. A set of protected tabulations based on 2010 Census responses, known as the 2010 Demonstration Data Products, were released in October 2019 to show data users how this new disclosure avoidance system might impact the accuracy of data products.

Data users gave feedback on the demonstration products to the Census Bureau both by email and at a workshop hosted by the National Academy of Sciences Committee on National Statistics in December 2019. Much of the feedback focused on concerns regarding the accuracy of the post-disclosure protected tabulations (i.e., how close the new tabulations were to the original tabulations) and bias (i.e., whether the new tabulations systematically differed from the original tabulations due to population size or other characteristics). Data users also highlighted specific geographies where accuracy was particularly important: counties, political entities such as incorporated places, and American Indian/Alaska Native/Native Hawaiian (AIANNH) Areas.

This document proposes a series of metrics to be used to assess the 2010 Demonstration Data Products as well as future development runs of the disclosure avoidance system (DAS) as improvements are made leading up to the release of 2020 Census data products. As testing and development of the disclosure avoidance system continues, these metrics will be used to concisely and quantitatively communicate data quality improvements to data users and the broader stakeholder community.

The intent is not to replicate a full analysis of each development run, but to provide a set of metrics that will inform stakeholders of the fitness of use across variables and geographies. Metrics will show the accuracy of both a broad set of demographic measures and specific types of use cases. The included metrics, and the formulation of metrics for specific use cases will evolve and new metrics will be added based on external feedback.

This document contains examples for the resident population of the United States. The resident population of Puerto Rico will be analyzed in a similar manner; however, statistics for the United States will not be pooled with statistics for Puerto Rico.

Metrics

Based on the feedback from the 2010 Demonstration Data Products, data users are concerned about accuracy, bias, and outliers.

Accuracy

Accuracy is measured by comparing the post-disclosure protected tabulations to the original, publically available tabulations from the 2010 Census and the internal pre-disclosure avoidance microdata from the 2010 Census.[1] Accuracy can be "absolute" or "relative" – that is, accuracy can be measured as either a count (the total population differed by 20 people) or as a percent of the original (the total population differed by 5%).

The following metrics for accuracy are proposed:

1. **Mean/Median Absolute Error (MAE)**: This is a measure of the "average" absolute value of the count difference for a particular statistic. For example, for total population at the county level, calculate Abs(MDF – CEF)[2] for each of the 3,143 counties, then take the median or mean.[3]
2. **Mean/Median Numeric Error (ME)**: This is a measure of the magnitude and direction of the average difference for a particular statistic. For example, for total population at the county level, calculate (MDF – CEF) for each of the 3,143 counties, then take the median or mean.
3. **Root Mean Squared Error (RMSE):** This is a measure of the square root of the average squared error for a particular statistics. It is the traditional measure of error for Census Bureau sample survey statistics. For example, for total population at the county level, calculate (MDF – CEF)^2 for each of the 3,143 counties, take the mean, then take the square root.
4. **Mean/Median Absolute Percent Error (MAPE)**: This is a measure of the "average" relative difference for a particular statistic. For example, for total population at the county level, calculate [Abs(MDF – CEF)/CEF] for each of the 3,143 counties, then take the median or mean.
5. **Coefficient of Variation (CV):** This is the relative error conterpart to RMSE. It is another traditional measure of error in Census Bureau sample survey statistics. For the same collection of statistics as was used for RMSE, calculate Avg(CEF), then calculate [RMSE/Avg(CEF)].
6. **Total Absolute Error of Shares (TAES):** This measure finds the proportion of each MDF value to the total MDF value for the summary geography and subtracts the proportion of the CEF value to the total CEF value for the summary geography. The absolute value of these proportional differences across evaluation geographies is then summed to the summary geography level. The goal is to provide a measure of the distributional error in the MDF shares.

---

[1] The post-disclosure protected tabulations are from the 2010 Demonstration Data Product Microdata Detail File (MDF) and subsequent runs of the disclosure avoidance system using differential privacy – referred to as "MDF." The publically available 2010 Census tabulations (post-swapping) are from the 2010 Census Hundred-percent Detail File (HDF). In order to make the results publically available, the initial analysis will be done based on the 2010 Census HDF tabulations, because these tabulations are already public via the 2010 Census Summary File 1. Internally, the Census Bureau will repeat this analysis using the 2010 Census Edited File (CEF) pre-swapped values.

[2] In this formula, and all the formulas that follow, MDF means "tabulated from the Microdata Detail File" and CEF means "tabulated from the Census Edited File." Most of the comparisons that the Census Bureau will present initially, and all of the comparisons that were done by external users of the 2010 Demonstration Data Products, substitute HDF for CEF in these formulas, meaning "tabulated from the Hundred-percent Detail File (swapped data)." The conceptually correct error measure is relative to the CEF, but in order to document the issues raised by external reviewers, the first collection of values for these metrics will be based on the HDF so that external users can verify that the Census Bureau has implemented the metric correctly. When subsequent versions of the 2020 DAS are used to generate new MDFs, they will be compared directly to the 2010 CEF.

[3] The reference to "counties" includes counties and county equivalents in the 2010 Census – the list of counties in the 2010 Census is located here: https://www.census.gov/geographies/reference-files/time-series/geo/tallies.html

7. **90th Percentile Absolute Percent Error**: This is a measure of the maximum likely error for the "bulk" of tabulated statistics (90 percent, following the U.S. Census Bureau Statistical Quality Standards).[4] For example, for total population at the county level, calculate [Abs(MDF – CEF)/CEF] for each of the 3,143 counties, then take the 90th percentile value. This will communicate to data users that, for the statistic in question, 90 percent of the post-disclosure protected statistics are within X percent of their 2010 Census internal pre-disclosure avoidance value.

Accuracy will be calculated using the above metrics both overall (e.g., for all 3,143 counties) and also for particular population and cell size categories (e.g., for counties with populations below 10,000 people or cells with counts equal to or greater than 100).

Bias

Bias is a concept related to accuracy, but direction of change and whether that varies with population size or other characteristics is what matters most. Prior research into the top-down algorithm (TDA) post-processing has demonstrated that geographic areas with small populations (or statistics with small cell sizes) tend to have a positive bias, where the privatized tabulation is systematically greater than the original tabulation, while those areas with larger populations (or larger cell sizes) tend to have a negative bias.

The following metrics for bias are proposed:

1. **Mean/Median Numeric Error (ME)**: This is a measure of the magnitude and direction of the average difference for a particular statistic. For example, for total population at the county level, calculate (MDF – CEF) for each of the 3,143 counties, then take the median or mean.
2. **Mean/Median Percent Error (MALPE)**: This is a measure of the magnitude and direction of the average relative difference for a particular statistic. For example, for total population at the county level, calculate [(MDF – CEF)/CEF] for each of the 3,143 counties, then take the median or mean.

Bias will generally be calculated by population size or cell size categories (e.g., categories for counties below 1,000 people, counties between 1,000 to 4,999 people, counties between 5,000 to 9,999 people, counties between 10,000 and 49,999 people, counties between 50,000 and 99,999 people, and counties equal to or greater than 100,000 people).[5] Bias will also be calculated by urban/rural classification and by percent non-Hispanic white population. Urban areas will be classified based on the Census Bureau's 2010 classification that require them to be comprised of densely settled core of census tracts and/or census blocks that meet minimum population density requirements, along with adjacent territory containing non-residential urban land uses as well as territory with low population density included to link outlying densely settled territory with the densely settled core.[6] "Rural areas" encompass all

[4] The Census Bureau's Statistical Quality Standards are available at: https://www.census.gov/content/dam/Census/about/about-the-bureau/policies_and_notices/quality/statistical-quality-standards/Quality_Standards.pdf

[5] Size categories will be evaluated to determine best fit and may be adjusted.

[6] To qualify as an urban area, the territory must encompass at least 2,500 people, at least 1,500 of which reside outside institutional group quarters. The Census Bureau identifies two types of urban areas: Urbanized Areas (UAs) of 50,000 or more people and Urban Clusters (UCs) of at least 2,500 and less than 50,000 people.

population, housing, and territory not included within an urban area. Using the metrics proposed above, the amount of bias introduced to urban and rural areas will be calculated.

Counties will be classified based on the percent of their population who were non-Hispanic white in the 2010 Census (e.g. counties with less than 10 percent population with 10-49 percent, and counties with 50 percent or more). This will provide insight into how the noise infused through the disclosure methodology is distributed across geographies with different racial and ethnic make-ups. The focus of these measures is to determine if the disclosure methodology has a tendency to either inflate or deflate the population by type of area or by characteristics of the population in an area.

For certain statistics and geographic areas, the distribution of proportional differences across subordinate geographies matters greatly. The metric **Total Absolute Error of Shares (TAES)** is proposed to measure how close the disclosure-protected spatial distribution is to the 2010 Census internal data distribution. It is calculated as follows: $\sum_i \left| \frac{MDF_i}{\sum_i MDF_i} - \frac{CEF_i}{\sum_i CEF_i} \right|$, where MDF_i is an individual subordinate geography's privatized tabulated value and CEF_i is an individual subordinate geography's 2010 Census value. To illustrate, imagine a county with two tracts: one that contains 90 percent of the county's population and one that contains the other 10 percent. If the privatized data now have equal populations in each tract for a hypothetical county, the TAES will be calculated as [Abs(0.5 - 0.9) + Abs(0.5 - 0.1)] = 0.8.

Outliers

Additionally, certain statistics and visualizations will be internally examined for "outliers": What is the largest increase in tabulated value? What is the largest decrease? These will inform internal evaluations about the plausibility of tabulated results. Since these outlier values may be connected to particular statistics and geographies, and could be used to back out private tabulated values, they are Title 13 restricted and will not be publically released. Counts of outliers will be made available externally, to allow for an assessment of the number of entities with exceptionally large differences from the original, private, tabulated statistics.[7]

Geographic Levels

Based on feedback received from the 2010 Demonstration Data Products, data users are particularly concerned about data fitness for states, counties, political entities such as incorporated places and minor civil divisions (MCDs), American Indian/Alaska Native/Native Hawaiian (AIANNH) Areas, and, for limited use cases, tracts and block groups. The first set of metrics will be produced for States, Counties, Places, and Tracts. Additional sets of metrics will be provided for Puerto Rico, as well as additional levels of geography such as MCDs, and AIANNH Area.

As changes are made to what is included in the "geographic spine" to improve accuracy across key geographies, measures may be provided for additional subsets, groups, or types of geographies.

Use Cases and Proposed Metrics

A general set of metrics were developed to provide an accuracy profile for a broad set of Census data – this accuracy profile will provide information on the fitness of use for many critical uses.

[7] Thresholds for what is considered an outlier will be determined based on use cases.

Additional metrics were developed for specific categories of use cases. Use cases were identified through a Federal Register Notice, the Committee on National Statistics (CNSTAT) Demonstration Products Workshop, and other outreach. Use case categories were created based on the type of accuracy that was the most important for the use cases within that category. While several measures of accuracy will be provided, each category has a primary measure for assessing fitness of use. This allowed for metrics to be developed that were designed specifically for the following categories of use cases:

Zero-Sum Total: Uses that rely on the accuracy of the distribution in addition to the overall accuracy because a fixed amount of something is being distributed across categories. For these uses, the accuracy needs may be greater for the distribution than for the actual estimates. For these types of use cases, the TAES would serve as the primary measure for fitness of use.

Zero-Sum Category: Same as zero-sum total except use cases rely on estimates for some subset of the total. For these types of use cases, the TAES would serve as the primary measure for fitness of use.

Variable-Sum Total: Similar to zero-sum use cases except that the total of what is being distributed can vary. For these types of uses, the accuracy of the estimate is more important than the accuracy of the distribution. For these types of use cases, the MAPE would serve as the primary measure for fitness of use.

Variable-Sum Category: Same as variable-sum total but for a subset of the population. For these types of use cases, the MAPE would serve as the primary measure for fitness of use.

Single Year of Age Accuracy: These use cases require accuracy for single years of age rather than age groups. For these types of use cases, the MAPE would serve as the primary measure for fitness of use.

Rates Accuracy: These uses cases rely on a measure of the size of a subgroup(s) within the total population. For these types of use cases, because they are based on a rate, the MAE and RMSE as a percentage point difference serves as the primary measure for fitness of use.

Percent Threshold: Use case depends on the subset of the population crossing a percent threshold. For these types of use cases, counts of entities crossing the threshold would serve as the primary measure for fitness of use.

Numeric Threshold: Use case depends on the subset of the population crossing a numeric threshold. For these types of use cases, counts of entities crossing the threshold would serve as the primary measure for fitness of use.

Basic Demographic Accuracy Profile

Total Population

Total population at the state level is invariant so a measure of accuracy is not needed. Measures will be provided for the county, place, and tract level. The county level includes counties and county equivalents. The place level includes incorporated places as well as census designated places. Additional sets of metrics will be provided for Puerto Rico, as well as additional levels of geography such as MCDs, and AIANNH Areas.

For the county and place level, the **MAE, RMSE, MAPE, CV, and MALPE** will serve as the primary measures of error. These will be produced by county and place size categories (less than 1,000 people, 1,000 to 4,999 people, 5,000 to 9,999 people, 10,000 to 49,999 people, 50,000 to 99,999 people, and equal to or greater than 100,000 people). The MAE and MAPE will serve as the primary measure of error, and the MALPE will serve as a measure of bias. [Tables 1 and 2]

Scatter plots of the distribution of errors for county and places will be produced for visual examination. [V1]

A secondary measure of outliers will be provided. This measure will include counts of counties and places where the absolute percent difference is "5 to 10 percent" and "above 10 percent" by size categories (less than 1,000 people, 1,000 to 4,999 people, 5,000 to 9,999 people, 10,000 to 49,999 people, 50,000 to 99,999 people, and equal to or greater than 100,000 people). [Tables 1 and 2]

For tracts, the primary error measures for total population, will be the **MAE and RMSE**. Because of the standard size of tracts, the tract-level measures will not be provided by size categories. A secondary measure will be provided for outliers, which will be the count of tracts where the absolute difference exceeds 10 percent. [Table 3]

For total population, additional measures of bias will be provided by urban and rural classification and by the percent of the population that is non-Hispanic white (<10%, 10% to 49%, and >50%). [Tables 4 and 5]

The urban/rural measure will be based on the block-level urban/rural designation. The block level **MAE, RMSE, MAPE, CV, and MALPE** for all urban blocks will be compared to the same measures for all rural blocks. [Table 4]

The non-Hispanic white measures will include the **MAE, RMSE, MAPE, CV, and MALPE** for counties by percent non-Hispanic white category (<10%, 10% to 49%, and >50%). [Table 5]

Total Housing Units
Counts of housing units are invariant at the block level; therefore a measure of accuracy is not needed.

Occupancy and Households
Measures will be provided for the county, place, and tract level. Because occupancy is expressed as a rate, the MAE, RMSE, and MALPE will be modified here to reflect the percentage point difference. The primary measure will be the modified **MAE, mean absolute percentage point error, and the modified ME, mean percentage point error** for the occupancy rate for counties and places. [Table 6] For tracts, the primary measure will be the modified **MAE, mean absolute percentage point error**. [Table 7]

A secondary measure will be counts for the county, place, and tract level, where the occupancy is 100 percent in the MDF but not the CEF, and where the occupancy is 0 percent in the MDF but not in the CEF. [Tables 6 and 7]

Review of the demonstration product revealed population, household size, and household counts that when considered together represented impossible values. This was due to inconsistencies between the

person file, which contains person information; and the housing unit file, which contains housing information that resulted from applying disclosure protections to each of these file separately. The following two measures are meant to show the extent of these inconsistencies. A count of tracts where households from the person file outnumber people when the count of people is derived from the household size variable will be provided. [Table 8] Even though the household size variable includes a "Size +7" category, by assuming those households all have the smallest size of 7, a population count can be obtained. This value can be compared to the population total from the person file. A count of the number of tracts where the population total is less than the population derived from the household size variable will also be provided. [Table 8]

The **MAE, RMSE and ME** for the persons-per-household derived by dividing the household population by the number of households will be provided for the county, place, and tract level. [Table 9]

Race and Hispanic Origin

The primary measure of accuracy for Hispanic origin and race for the state, county, and place level will be the **MAE, RMSE, MAPE, CV, and MALPE**. Measures will be produced by all states, counties, and places, as well as county and place size categories (by counties and places with between 0 and 9 people, between 10 and 99 people, and equal to or greater than 100 people of the race/Hispanic origin category. The **MAE and RMSE** will be used at the tract level for the same Hispanic origin and race categories.

Error measures will be provided in a table by the following Hispanic origin and race groupings:

- Hispanic or Latino Origin [Tables 10, 11, 12, and 13]
- 6 Major Race Groups Alone (White, Black, American Indian and Alaska Native (AIAN), Asian, Native Hawaiian or Other Pacific Islander (NHPI), and Some Other Race (SOR)) and a Two or More Category by Hispanic or Latino Origin [Tables 14.a-g, 15.a-g, 16.a-g, and 17.a-g]
- 6 Major Race Groups Alone or In Combination (White, Black, AIAN, Asian, NHPI, and SOR) by Hispanic, not Hispanic or Latino Origin [Tables 18.a-f, 19.a-f, 20.a-f, and 21.a-f]
- Number of Races Groupings – one race, two races, three races, four races, five races, and six races [Tables 22.a-f, 23.a-f, 24.a-f, and 25.a-f]

To supplement analyses conducted by other areas for the redistricting data product, we will also create the following Hispanic origin and race groupings by voting-age population (18 years and older) at the tract and block group levels:

- 6 Major Race Groups Alone (White, Black, AIAN, Asian, NHPI, and SOR) and a Two or More Races Category by Hispanic or Latino Origin for the Population 18 and Over [Tables 26.a-g and 27.a-g]
- 6 Major Race Groups Alone or In Combination (White, Black, AIAN, Asian, NHPI, and SOR) by Hispanic or Latino Origin for the Population 18 and Over [Tables 28.a-f and 29.a-f]
- Hispanic or Latino Origin by number of race groupings for the Population 18 and Over [Tables 30.a-f and 31.a-f]

Age and Sex

The primary measures of accuracy for age and sex will be the **MAE, RMSE, MAPE,CV, and MALPE**. These will be produced for the county and place geographic levels.

Error measures will be provided for the following sex by age groupings:

- Ages 0-17, 18-64, and 65 and over [Tables 32 and 33]
- Age in 5-year age bins from 0-115 [Tables 34 and 35]

Population pyramids will be produced for counties representative of the five size categories for visual examination. [V2]

Group Quarters Population by Major GQ Type and Institutionalized versus Noninstitutionalized
The primary measure of accuracy for group quarters type will be the **MAE, RMSE, MAPE,CV, and MALPE**. These will be produced at the county and place level for the seven major group quarters types and for the institutionalized and noninstitutionalized population for the following total population size categories: less than 1,000 people, 1,000 to 4,999 people, 5,000 to 9,999 people, 10,000 to 49,999 people, 50,000 to 99,999 people, and equal to or greater than 100,000 people. The **MAE and RMSE** will be used at the tract level for the same GQ categories. [Tables 36, 37, and 38]

Major GQ Types are classified as:

- **Institutional Group Quarters:** 1) Correctional Facilities for Adults, 2) Juvenile Facilities, 3) Nursing Facilities/Skilled-Nursing Facilities, 4) Other Institutional Facilities
- **Noninstitutional Group Quarters**: 5) College/University Student Housing, 6) Military Quarters, 7) Other Noninstitutional Facilities

Categories of Use Cases with Specific Examples
Emergency Service Planning for a Specific Population within a Small Geographic Area
Variable-sum category (local)
A specific example of this type of use case is a scenario where the number of people aged 75 and over is required to determine the number of buses or other resources needed to evacuate the elderly population from an area. This type of use case is representative of a local, non-zero-sum category use case since the number of buses is not limited and will be based on the size of the population in need. This makes the size of the target population the population measure that requires accuracy. There is also a geographic need, since the buses would need to be staged in close vicinity to the population in need. This type of use case tends to be for smaller geographic areas and most often requires counts of the elderly or of children.

The primary selected measure that will be provided as an indication of the fitness of use of data for this use case is the **MAE and RMSE** at the tract level for the population aged 75 and over. [Use Case Table 1]

Counts of the tracts that exceed a numeric difference of 10 percent for the target population group will be provided as a secondary measure of fitness for use. [Use Case Table 1]

Shaded tract-level maps of the absolute difference for the population aged 75 and over will be provided for visual examination. [Use Case Visualization 1]

These measures will be repeated for young children (under 5 years of age) and other age groups based on external input. [Use Case Table 2]

Distribution of Federal Funds

Zero-Sum Total

The distribution of federal funds use case is generally understood to be a state, county, and place level distribution of a fixed amount. Because state-level counts are invariant, a state level measure isn't needed. With this type of use case, a fixed amount is distributed based on each area's share of the population, making the accuracy of the shares, or the distribution, the primary measure that requires accuracy.

The primary measure to assess fitness of use for this use case will be the **TAES** at the county level as a share of the nation, at the county level within each state as a share of that state, and at the place level as a share of that state. [Use Case Tables 3, 4, and 5]

Projections of the Population Entering School or Eligibility for a Program

Single Year of Age Accuracy

This use case requires accuracy for counts of people of a single year of age or age ranges. For this type of use case, single year of age accuracy may be needed for a single year of age or for an age range, for example, those entering school, or those who will be graduating school, or those who will be eligible for different programs for a set number of years in the future. Other examples include those expected to complete immunization schedules; expected draft registration; eligibility for retirement, Medicare, or Social Security; or, more broadly, single year of age projections.

These use cases include the county, place, and tract levels of geography. The measures of accuracy for assessing fitness of use for this use case are the same as for the total population, but applied to a specific age or age range. The accuracy need is in the counts of the population in the specified age or age range.

For the county and place level, when a single age, or age range is being considered, the **MAE, RMSE, MAPE, CV, and MALPE** will be the primary measure of error. These will be produced for ages 4 and 17, by county and place size categories (less than 1,000 people, 1,000 to 4,999 people, 5,000 to 9,999 people, 10,000 to 49,999 people, 50,000 to 99,999 people, and equal to or greater than 100,000 people. [Use Case Table 6, 7, 8, and 9]

Initially, this measure will be produced for ages 4 and 17, with additional ages and age groups to be added based on external input. An aggregate measure of single year of age accuracy, the **TAES**, will also be provided for the share of counties and places within the nation. [Use Case Table 10, 11]

Population pyramids for selected counties and places will be provided for visual examination. [Use Case Visualization 2, 3]

For tracts, the primary measure will be the **MAE and RMSE**. A secondary measure will be provided for outliers, which will be the count of tracts where the absolute numeric difference exceeds 10 percent, or some other threshold from stakeholder input, of the average size across tracts for that age or age range. [Use Case Table 12, 13]

Total Population for American Indian and Alaska Native Race Groups

Zero- and Variable-Sum Category

Federal funding use allocation formulas such as the Tribal Transportation Programs and Indian Housing Block Grant funding rely on Census data. These uses require accuracy of the counts of the American Indian and Alaska Native population.

The measures of accuracy used for this use case will be similar to those used for Hispanic origin and race groups, with the addition of the **TAES** measure applied to this specific race group.

The **MAE, RMSE, MAPE, CV, and MALPE** will be used at the state, county, and place level for the AIAN population alone and in combination. [Use Case Table 14]

The **TAES** measure will be applied to the AIAN population distribution across counties and places within the nation. [Use Case Table 15]

Outreach for Rare/Small Populations – Race Use Cases

Variable-Sum Total

This use case depends on the accuracy of the data for locating rare or small populations – these measures will focus on how accurately the presence of AIAN and NHPI alone populations can be determined. Fitness of use depends on being able to correctly identify the target population with a minimal number of false positives or false negatives, or the ability to show when a population exists in an area, and when it does not exist.

The primary measure of accuracy for rare and small populations (AIAN and NHPI alone) for the county and place level will be the **MAE, RMSE, MAPE, CV, and MALPE**. Measures will be produced by county and place size categories (by counties and places with between 0 and 9 people, between 10 and 99 people, and equal to or greater than 100 people in the small population group (AIAN and NHPI alone)). [Use Case Table 16, 17, 18, and 19]

For tracts, the primary measure will be the **MAE and RMSE**. [Use Case Table 20 and 21]

A secondary measure of fitness for use will be provided to identify clusters of AIAN and NHPI population, with the minimum population to indicate a cluster being the presence of at least 100 people in a tract that are either AIAN (alone or in combination) or NHPI (alone or in combination). A count of false negatives and false positives will be provided for tracts. A false positive will be defined as when the CEF population is equal to or greater than 100 and the MDF population is less than 20. A false negative will be defined as when the CEF population is less than 20 and the MDF population is equal to or greater than 100. [Use Case Table 20 and 21]

Target Vacancy/Occupancy Rates

Percent/Rate Thresholds

In this use case, a threshold has been established as a target or as a threshold for inclusion. A specific example is the use of vacancy rates as an indication of housing availability.

To obtain a measure of fitness for use for this use case example, counts of counties, places, and tracts where the occupancy rate exceeds 90 percent in the MDF, but is below 90 percent in the CEF will be provided. [Use Case Table 22]

Counts for other thresholds will be added based on external input.

Additional Funding for Public Services

Numeric Thresholds

In this use case, a threshold has been established where once an area crosses that threshold, additional funds to meet the needs of the area are made available. A specific example is the provision of additional funds to hire additional police officers once an area exceeds a population of 50,000.

To obtain a measure of fitness for use for this use case example, counts of counties, place level geographies, and tracts where the population exceeds 50,000 in the MDF, but is below 50,000 in the CEF, and where the population is below 50,000 in the MDF, but below 50,000 in the CEF will be provided. [Use Case Table 23]

Counts for other thresholds will be added based on external input.

<u>Full Demographic and Housing Characteristics File (DHC) Variables Use Cases</u>

The Tenure and Relationship variables, planned for inclusion in the DHC, were not available in the 2010 Demonstration Data Products and will not be available until the DAS is fully scaled up. Metrics for these variables have been developed and will be shared at a future meeting.

Appendix: Measures of Accuracy

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percent Error (MAPE)
- Coefficient of Variation (CV)
- Mean Algebraic Percent Error (MALPE)
- Root Mean Squared Error
- Percent Difference Thresholds
- Total Absolute Error of Shares

Mean Absolute Error (MAE) = (Σ (|MDF – CEF|))/N
MAE takes the absolute value of the difference between the MDF and the CEF value for each evaluation geography, sums them, and divides by the number of evaluation geographies. The goal is to provide an easy to interpret measure of the numeric error.

Root Mean Squared Error = SQRT(Σ ((MDF – CEF)^2)/N)
This measure squares the difference between the MDF and the CEF number for each evaluation geography, sums these values across evaluation geographies, divides by the number of evaluation geographies, and finds the square root of this value. It presents an alternative measure that places greater emphasis on large numeric errors versus mean absolute errors.

Mean Absolute Percent Error (MAPE) = ((Σ ((|MDF – CEF|)/ CEF))/N)*100
MAPE takes the absolute value of the difference between the MDF and the CEF value for each evaluation geography, divides that by each respective CEF value, sums them, divides by the number of evaluation geographies, and multiplies the result by 100. The goal is to provide an easy to interpret relative measure of error. This is one of the most commonly used measures for assessing the accuracy of a series of population estimates.

Coefficient of Variation = (RMSE/(Σ (CEF)/N))*100
This measure restates the RMSE as a percentage of the average statistic in the geography.

Mean Algebraic Percent Error (MALPE) = ((Σ((MDF – CEF)/CEF))/N)*100
MALPE takes the difference between the MDF and the CEF value for each evaluation geography, divides that by each respective census value, sums them, divides by the number of evaluation geographies, and multiplies the result by 100. Its purpose is to identify systematic bias and provide an alternative for a relative measure of error.

Percent Difference Thresholds = Number of percent differences above a certain threshold
Unlike the other measures, Percent Difference Thresholds is a numeric value that relies upon an arbitrarily set threshold (e.g., 5 and 10 percent). In short, the percent difference is computed by dividing the difference between the MDF and CEF value for a given area by the CEF value for that area and multiplying by 100. The end measure simply represents a count of how many evaluation geographies in the summary area exceeded a particular threshold in their absolute percent difference of the estimate. It provides an intuitive measure of the distribution of differences.

Total Absolute Error of Shares = Σ|((MDF/ΣMDF) – (CEF/ΣCEF))|
This measure finds the proportion of each MDF value to the total MDF value for the summary geography and subtracts the proportion of the CEF value to the total CEF value for the summary geography. The absolute value of these proportional differences across evaluation geographies is then summed to the summary geography level. The goal is to provide a measure of the distributional error in the MDF shares.